Supporting Information Appendix

Atomic-level characterization of protein-protein association

MD Simulation Details

Nonbonded interactions were truncated and shifted at a cutoff distance, r_{cut}, and long-range electrostatics were computed in *k*-space using a grid-based method (Gaussian spreading to the grid was used[1]). To further increase computational efficiency in the soluble-protein simulations, we modified all the hydrogen, H, and water oxygen, O_w, masses (H: 4 Da; O_w: 10 Da); this allowed us to integrate the equations of motion using a RESPA scheme[2] with an inner time step of 4.5 fs and an outer time step of 9 fs.[3] Previous simulations of a well-characterized protein, villin, have shown that using a larger time step and altering the masses of hydrogen and water oxygens do not have any substantial effect on its kinetics or the thermodynamics.[4] The simulations of CLC-ec1 did not have modified masses and used an inner time step of 2.5 fs and an outer time step of 7.5 fs.

Each prepared system was minimized using Desmond/GPU[5] and then equilibrated in the NPT ensemble for 50 ns with harmonic position restraints on all heavy protein and ligand atoms, tapered linearly to 0 from 5 kcal mol^{-1} $Å^{-2}$. For the soluble proteins, production runs were subsequently performed in the NVT ensemble from the final frame of the NPT relaxation simulation at 300 K. For CLC-ec1, production runs were performed in the NPT ensemble at 310 K. Equilibration and production simulations were all performed on Anton, a special-purpose machine capable of running very long MD simulations.[13]

Simulations of barnase-barstar (BB_0_0, BB_0_1, BB_0_2, and condition BB_1) contained 4 Na^+ atoms in a cubic box of length 82 Å with approximately 56,000 atoms. Simulation conditions for barnase-barstar in conditions BB_2 and BB_3 were similar, except that the TIP4P/2005 water model was used. Residues 40 and 82 in barstar were mutated to cysteines such that its sequence was consistent with the wild-type protein. All barnase-barstar simulations used an r_{cut} of 10 Å. Torsional backbone corrections with k = 4 kcal mol^{-1} were applied to residues 5–108 on barnase and residues 3–87 on barstar for BB_0_0, BB_0_1, BB_0_2, and condition BB_1, and with k = 1 kcal mol^{-1} for condition BB_2. No torsional corrections were applied in condition BB_3.

Insulin dimer simulations (IND_0_0, IND_0_1, IND_0_2, IND_1_0, IND_1_1, IND_1_2, and condition IND_2) contained 4 Na^+ atoms in a cubic box of length 69 Å with approximately 32,000 atoms, and used an r_{cut} of 10 Å. Torsional backbone corrections with k = 5 kcal mol^{-1} were applied to all backbone residues. Ras–Raf-RBD simulations (RAS_0_0, RAS_0_1, RAS_0_2, and condition RAS_1) contained 4 Na^+ atoms, 1 Mg^{2+} atom, and a GppNHp molecule bound to Ras, in a cubic box of length 88 Å with approximately 68,000 atoms, and used an r_{cut} of 10 Å. Torsional backbone corrections with k = 1 kcal mol^{-1} were applied to all backbone residues. His^{166} in Ras was simulated in its protonated state. Simulations of RNaseHI–SSB-Ct (RNA_0_0, RNA_0_1, RNA_0_2, and condition RNA_1) contained 2 Na^+ atoms in a cubic box of length 99 Å with approximately 96,000 atoms, and used an r_{cut} of 13.5 Å. Torsional backbone corrections with k = 1 kcal mol^{-1} were applied to RNase-HI only; the SSB-Ct peptide is disordered in solution and folds upon binding.

Simulations of TYK2-pseudokinase (TYK_0_0, TYK_0_1, TYK_0_2, TYK_1_0, TYK_1_1, TYK_1_2, and condition TYK_2) contained 3 Na^+ atoms and 2 molecules of compound 7012,

one bound to TYK2 and one bound to its pseudokinase, in a cubic box of length 105 Å with approximately 115,000 atoms, and used an r_{cut} of 10.0 Å. Torsional backbone corrections were applied to TYK2 with k = 2 kcal mol^{-1} and to the pseudokinase with k = 1.5 kcal mol^{-1}. Residues in loops that were not resolved in the crystal structure were not included in the simulations, and all chain termini were capped with neutral groups. His^{669} in the pseudokinase was simulated in its protonated state.

Finally, simulations of CLC-ec1 dimer (CLC_0, CLC_1) had an NaCl concentration of 150 mM in a 120 Å × 120 Å × 107 Å box with approximately 156,000 atoms, and used an r_{cut} of 9.166 Å. Torsional backbone corrections were applied to CLC with k = 1 kcal mol^{-1}. Glu^{113} in chains A and B, and Asp^{417} in chain A, were simulated in their protonated states.

As in previous MD simulation studies of spontaneous protein-protein association,[6,7,8,9] the starting monomer conformations were taken from crystal structures of the associated protein-protein complexes. Such a choice might appear to "help" the simulations find the native-complex structure, similar to the way it can often improve the observed accuracy in protein-protein docking studies. We do not believe that, in practice, this choice leads to any bias in our simulations, for two reasons: First, the MD simulations explicitly model protein atomic motions and allow for the protein monomers to deviate, sometimes significantly, from their initial structures. Second, for the systems studied in this work, the observed experimental difference between protein monomers in the bound and the unbound complex is small (SI Appendix Table S1).

Justification for torsional corrections

Applications of torsional corrections for these protein-protein systems can be justified in the following three ways: (i) For all five protein-protein complexes simulated in this study, the interface Cα RMSD of the experimentally determined monomer structures differed very little from their experimentally determined complex structures (SI Appendix Table S1), implying that formation of the complexes does not require large conformational changes in the monomers (in particular, for protein-peptide systems like RNaseHI–SSB-Ct, a structure-based analysis of existing protein-peptide structures suggests that most peptides do not induce conformational changes in their binding partners when they associate[10]); (ii) where known, the experimental values of the on-rates for these complexes are, in the context of protein-protein association, diffusion-limited or faster than diffusion-limited, meaning that experimental association timescales are likely too fast to allow for large conformational changes upon binding; (iii) for barnase-barstar, we ran simulations with and without torsional corrections, and the simulations without torsional corrections (condition BB_3) gave an estimate of the association rate, $k_{on} = 6.5 \times 10^6\ M^{-1}\ s^{-1}$, that differed an order of magnitude more from the experimentally determined association rate, $6.0 \times 10^8\ M^{-1}\ s^{-1}$, than did the estimates of the simulations performed with the torsional corrections ($5.8 \times 10^7\ M^{-1}\ s^{-1}$ for condition BB_2; see below for details about how these rates were calculated). In other words, we observed fewer binding events per unit time in our simulations than would be expected based on the experimentally known rate; simulations with torsional corrections, however, yielded a rate that was approximately an order magnitude faster than the estimate derived from the simulations without torsional corrections, and closer to the experimental value.

Using the Amber force field in combination with torsional corrections led to tempered binding simulations yielding accurate predictions of native-complex structures for the systems studied in this work. We believe that this approach will often (but not always) work for other systems in which the individual monomers also do not undergo large conformational changes upon binding. In systems in which the conformations of the monomers differ significantly from their conformations in the bound state, however, torsional corrections are not applicable, and it is likely that modern force fields will have more difficulty modeling such systems.

Tempered Binding Simulation Details

In practice, our tempered binding simulations had 120 rungs, subdivided into 3 overlapping regions of 60 rungs each, and a separate simulation, or walker, was initiated in each of the three regions. In order to reduce computational costs during an update of λ_i, a walker was only allowed to transition within its own subdivided region. Walkers could exchange between these subdivided regions of rung space whenever two walkers jointly occupied an overlapping region. The free energy weights at rung i, f_i, were calculated adaptively during the simulation using an approach similar to the one described in ref. 11.

When van der Waals interactions were scaled, we used a softcore potential of the form:[12]

$$V_{\mathrm{softcore,i}}(r) = 4\epsilon\lambda_i \left[\left(\frac{1}{\alpha(1-\lambda_i)^2+r^6/\sigma^6}\right)^2 - \frac{1}{\alpha(1-\lambda_i)^2+r^6/\sigma^6}\right].$$

Here, r is the distance between two atoms, $\alpha = 0.5$, and ϵ and σ are the Lennard-Jones interaction strength and radius. Near electrostatic energies were linearly scaled.

In tempered binding simulations, interactions should be scaled enough that dissociation from bound complexes happens relatively quickly on simulation timescales, but not so much that protein association timescales become slow. The optimal value of λ at the top rung should thus depend on both the expected binding affinity of the sampled complexes—complexes with higher binding affinities will require more scaling, whereas complexes with lower binding affinities will require less—and the effect of the scaling on the association rate of the two protein monomers. The scaling values used for this work were determined with these observations in mind; there is ongoing work toward a more systematic approach to the optimal choice of λ at the top rung. We note that, although the amount of scaling affects the sampling efficiency, it does not bias the sampling toward any particular state (e.g., the native complex). In other words, the distribution of states at rung 0 will be the same regardless of the amount of scaling (although the simulation time required to converge such a distribution may be much greater when the choice of scaling is non-optimal).

In BB_0, at the top rung of the ladder, the electrostatic and van der Waals interactions between barnase and barstar were scaled by 0.833 and 0.867, respectively; the update interval was 10.8 ps. For the other tempered binding simulations, we used a "split charge" tempering approach: The atoms in the protein monomers (which will here be labeled A and B) were split into groups that contained only atoms with a positive (negative) partial charge, A_{pos} and B_{pos} (A_{neg} and B_{neg}), and the interactions between positive and negative groups, A_{pos}-B_{neg} and A_{neg}-B_{pos}, were separately scaled. In IND_0, split charge scaling down to 0.9 and 0.995 was used between the insulin monomers and between each protein monomer and all the water molecules, respectively. In RAS_0, split charge scaling down to 0.99 was used between Ras and GppNHp and the Raf-Ras binding domain. In RNA_0, split charge scaling down to 0.992 was used between RNaseHI and the disordered SSB-Ct peptide. Finally, in TYK_0 and TYK_1, split

charge scaling down to 0.985 was used between TYK2 and its bound compound 7012, and the pseudokinase and its bound compound 7012. For CLC_0 (CLC_1), split charge scaling down to 0.995 (0.925) and 0.990 (0.980) was used between CLC monomers and between each protein monomer and the lipid bilayer, respectively. The update interval was 5.4 ps for IND_0, RNA_0, TYK_0, and TYK_1, and 4.8 ps for CLC_0 and CLC_1.

In the tempered binding calculations performed for this work, several hundred microseconds of simulated chemical time were required to observe reversible binding. The systems simulated for this work ranged in size from 32,000 to 156,000 atoms, and were typically run on a 64-node Anton 2 machine (on which simulations of a few hundred microseconds of chemical time take on the order of 1–2 weeks of machine time.)[13] There was only a small slowdown (~15–20%) associated with the tempered binding simulations in this work compared to conventional molecular dynamics simulations on Anton 2. Because Anton 2 can efficiently simulate systems as large as one million atoms,[13] we believe that tempered binding can be realistically applied to much larger systems than those simulated for this work. (We note that the atom count referred to here is a count of the all the atoms in the system, including the proteins, water, ions, and lipids, if needed.)

Representative Cluster Analysis

To obtain a representative structure without bias for Fig. 1, we used the following clustering approach. For each protein-protein system, an all-to-all backbone RMSD[14] was calculated over frames from rung 0 of the tempered binding trajectories. A hierarchical clustering approach[15] was then used to cluster the frames using this all-to-all RMSD matrix. The simulation frames

chosen for Fig. 1 are representative frames[16] from the most occupied cluster. VMD[17] was used to visualize trajectories, and molecular images were rendered using PyMol.[18]

Estimating the Barnase-Barstar Free Energy of Association

We estimated the free energy of binding for barnase-barstar in two ways: (1) directly from the tempered binding simulations (BB_0) and (2) by running additional biased tempering simulations for better statistics.

In BB_0, an intermediate rung was determined where multiple reversible transitions between associated and unbound states were observed. The free energy cost for the associated state to transition from rung 0 to this intermediate rung was calculated using the Bennett Acceptance Ratio (BAR) method,[19] and then the reversible free energy of association was calculated at the intermediate rung. To complete the thermodynamic cycle, separate simulations of barnase and barstar were performed with protein-water tempering to account for the free energy of scaling the protein-water interaction. The sum of all these contributions gave −21 kcal mol^{-1}.

To increase the statistical efficiency of our calculation of the binding free energy, we used restraints in our tempering procedure that employ information about the native complex, a method similar to the umbrella sampling protocol described in Gumbart et al.[20] A flat-bottom harmonic well (FBHW) restraint with the form

$$FBHW(x) = \begin{cases} \frac{1}{2}k(x - eq_0)^2, & x < eq_0 \\ 0, & eq_0 \leq x \leq eq_1, \\ \frac{1}{2}k(x - eq_1)^2, & x > eq_1 \end{cases}$$

with constants eq_0 = 0 Å, eq_1 = 5.514 Å, and k = 25 kcal mol^{-1} Å^{-2}, was added between the centers of mass of the barnase and barstar interfaces (defined by heavy atoms within 4.5 Å of each other in the crystal complex). This restraint was chosen so that bound configurations would be unlikely to interact with the walls of the restraint. Over the next 100 rungs, the near electrostatics between the two proteins, excluding interactions between the atoms at the interface, were scaled linearly to 95% of their original strength using a split charge approach. In addition, energy terms for, and interactions of, the interface atoms were scaled linearly such that the 100th rung was equivalent to simulating with the replica exchange with solute tempering 2 (REST2) approach[21] at a temperature of 533 K, where the solute was defined as any interface atom. Note that for pairs of atoms involved in both split-charge and REST2 scaling, the scale factors were multiplied. From the 101st rung to the 900th rung, the flat-bottom harmonic restraint's parameters, eq_0 and eq_1, were scaled linearly to 86.6 Å, allowing the protein-protein system to associate and dissociate reversibly. The final 100 rungs scaled all interactions back to their values at rung 0, while keeping the FBHW fixed at 86.6 Å. Three sets of three simulations, where exchange was permitted within each set of simulations, were run with this protocol, each to 45 μs, in a cubic box of edge length 100 Å. BAR was performed to calculate the free energy difference between the first and last rungs. The last window was unbiased to the unbound state numerically by accounting for the effect of the flat-bottom harmonic restraint.[20] The free energy was found to be −19.2(2) kcal mol^{-1}, with errors calculated as a standard deviation over the three sets of simulations.

Estimating the Association Rates From Conventional MD Simulations

To estimate the k_{on} of barnase-barstar association from simulation, we first pooled all 61 conventional association simulations under condition BB_1 of Table 1, and calculated T_u, the

total amount of time the proteins are not bound in the native complex (defined as times before which the interface Cα RMSD to the native complex, after a 10-ns moving average, decreased to 1.5 Å). Assuming pseudo-first-order association kinetics, association events can be modeled as a Poisson process sampled over an interval of length T_u. Aggregating across the 61 simulations under condition BB_1 gives T_u = 337.61 μs. A total of 24 association events were observed in these simulations. The maximum likelihood estimate of the rate of a Poisson process is the number of observed events divided by the time period of observation. For a given number of proteins, the protein concentration, and thus the rate of the Poisson process under observation, will scale inversely with the number of water molecules in the system. The maximum likelihood estimate for k_{on} evaluates to 2.3(2) × 10^7 M^{-1} s^{-1} at 300 K. To calculate the error, trajectories were blocked into groups with aggregate simulation times of approximately 85 μs. The error is a standard error of the mean calculated over those five blocks. The error bars for the rate in BB_2 were calculated with blocks of approximately 40 μs. The dissociation rate was calculated as the product of the association rate, k_{on}, and dissociation constant, K_d. The association rates for other systems were calculated in a similar way and are listed in Table 1.

Schreiber and Fersht[22] measured k_{on} = 6.0 × 10^8 M^{-1} s^{-1} at 298 K for barnase-barstar. Increasing the salt concentration reduced the rate of association due to electrostatic screening effects. In the presence of 100 mM and 500 mM NaCl, k_{on} dropped to 1.1 × 10^8 M^{-1} s^{-1} and 1.6 × 10^7 M^{-1} s^{-1}, respectively. The ionic concentration in our simulations was low (~10 mM), so the simulation-derived rate should be compared to 6.0 × 10^8 M^{-1} s^{-1}, which implies that the association of barnase-barstar in simulation is about 25-fold too slow. The slower kinetics are consistent with the observation that simulations of proteins in TIP3P tend to favor collapsed and aggregated states.[23,24] The k_{on} estimate for simulations done in a more accurate water model, TIP4P/2005, is faster (5.8 × 10^7 M^{-1} s^{-1}, with 16 events, BB_2) and is only 10-fold slower than the experimental

value. In TIP3P, the k_{on} estimate for simulations done without any torsional corrections (4.4×10^6 M^{-1} s^{-1}, with 3 events, BB_3) is 140-fold too slow, illustrating the importance of the torsional corrections for a proper description of the association process.

Transition Path Analysis

Encounter-complex analysis

Encounter complexes were extracted from the spontaneous-association trajectories in Table 1, and defined as simulation snapshots between the time when the two proteins first made contact and when they either dissociated (an unsuccessful transition pathway) or formed the native complex (a successful transition pathway). Unsuccessful encounter complexes with lifetimes shorter than 100 ns were not included in the analysis. If a simulation formed a complex that never relaxed into the native complex or dissociated during the course of the simulation, it was also not included in the analysis. Simulation frames were uniformly sampled from encounter complexes and aligned to the larger protein.

Several of our trajectories remained trapped in metastable states that never dissociated nor found the native complex during the course of the simulation, so it is possible that at longer timescales (>100s of μs), more varied dynamical behavior (e.g., more protein-protein interface exploration before binding) could be observed. The fact that our predicted kinetics for these protein-protein systems are in the right range for diffusion-limited complexes (Table 1), however, suggests that our simulation timescales and sampling are sufficient for capturing the association mechanism. In particular, our estimates for the association rate of barnase-barstar are slower than

experimental association timescales, indicating that these long-lived metastable traps might be due to inaccuracies in the force field (e.g., due to the TIP3P water model favoring collapsed protein states, as discussed above).

The spheres in Fig. 3 correspond to the position of the Cα atom of a centrally located binding-interface residue in the smaller protein: Asp^{39} on barstar, Phe^{29} on insulin, Arg^{89} on Raf-RBD, Phe^{177} on SSB-Ct, and Phe^{672} on the pseudokinase. The angles used in SI Appendix Fig. S5 to describe the orientation of the smaller protein relative to the larger protein were defined following Gumbart, Roux, and Chipot.[20] The centers of mass of three points on the larger protein (P1, P2, and P3) and three points on the smaller protein (P1', P2', P3') were defined, and the Euler angles, Θ (P1−P1′−P2′), Φ (P1−P1′−P2′−P3′), and Ψ (P2−P1−P1′−P2′), were calculated. The selections used are given in SI Appendix Table S4. Since we were mainly concerned with the orientation of the smaller monomer, the P3 point was never used, but is still included in SI Appendix Table S4 for completeness.

Probability of association

We computed the probability of association for a given simulation snapshot by initializing between 20 to 100 additional simulations from that snapshot. For each of these additional simulations, the initial atom positions corresponded to those of the chosen snapshot, but random velocities for each atom were drawn from a Boltzmann distribution. The lengths of these additional trajectories, which we term *shots*, were mostly on the order of 1–10 μs, although some were as long as 60 μs. The probability of association (or committor value), p_{Assoc}, for each snapshot was computed as the fraction of shots initiated from that snapshot in which the proteins associated before unbinding. The error on the committor was calculated as:[25]

$$\sqrt{\frac{p_{\mathrm{Assoc}}(1-p_{\mathrm{Assoc}})}{N}},$$

where N is the total number of shots. A trajectory was considered bound when the interface Cα RMSD to the native complex, after a 10-ns moving average, decreased to 1.5 Å. A trajectory was considered unbound when the minimum distance between every pair of heavy atoms of the two monomers, after a 10-ns moving average, exceeded 4.5 Å. For some committor attempts, the shot ended with neither an unbound complex nor a native complex, and were thus labeled "inconclusive."[26] We did not consider committor values for configurations where more than half of the shots were inconclusive.

Fraction of native contacts and fractional hydration

The fraction of native contacts was calculated using the formula from Best, Hummer, and Eaton (first equation in the supporting information of ref. 27), with $\beta = 5$ Å^{-1} and $\lambda = 2$. Native contacts were defined as pairs of heavy atoms in the crystal structure of the complex within 4.5 Å of each other, one from each monomer.

The fractional hydration, w, was defined such that its value was around 1 when the native binding interface between the two proteins was solvated to a level similar to that when the proteins were unbound, and defined such that its value was around 0 when the interface was solvated to a level similar to when they were bound in the native structure. For each system, we determined the average number of water molecules, $\langle N_w \rangle$, near the binding interface of each protein (e.g., A and B) when they were bound and unbound in conventional MD trajectories. The fractional hydration was then calculated as follows:

$$\langle N_{\mathrm{w}}^{\mathrm{bound}} \rangle = \frac{1}{2}(\langle N_{\mathrm{w,A}}^{\mathrm{bound}} \rangle + \langle N_{\mathrm{w,B}}^{\mathrm{bound}} \rangle)$$

$$\langle N_{\mathrm{w}}^{\mathrm{unbound}} \rangle = \frac{1}{2}(\langle N_{\mathrm{w,A}}^{\mathrm{unbound}} \rangle + \langle N_{\mathrm{w,B}}^{\mathrm{unbound}} \rangle)$$

$$N_{\mathrm{w}}(t) = \frac{1}{2}(N_{\mathrm{w,A}}(t) + N_{\mathrm{w,B}}(t))$$

$$w(t) = \frac{N_{\mathrm{w}}(t) - \langle N_{\mathrm{w}}^{\mathrm{bound}} \rangle}{\langle N_{\mathrm{w}}^{\mathrm{unbound}} \rangle - \langle N_{\mathrm{w}}^{\mathrm{bound}} \rangle}$$

A water molecule was considered to be near the binding interface if its oxygen atom was within 4 Å of a protein heavy atom at the native binding interface. Here, the native binding interface was defined as all heavy atoms in one monomer within 4 Å of the heavy atoms in the other monomer to focus the selection on waters that are directly interacting with the interface.

Supporting References

1. Shan, Y., Klepeis, J. L., Eastwood, M. P., Dror, R. O. & Shaw, D. E. Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J. Chem. Phys.* **122(5),** 054101 (2005).

2. Tuckerman, M., Berne, B. J. & Martyna, G. J. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.* **97(3),** 1990–2001 (1992).

3. Feenstra, K. A., Hess, B. & Berendsen, H. J. C. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **20(8),** 786–798 (1999).

4. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. U. S. A.* **110(15),** 5915–5920 (2013).

5. Bergdorf, M., Baxter, S., Rendleman, C. A. & Shaw, D. E. Desmond/GPU performance as of November 2016. D. E. Shaw Research Technical Report DESRES/TR--2016-01 (2016).

6. Ahmad, M., Gu, W. & Helms, V. Mechanism of fast peptide recognition by SH3 domains. *Angew. Chem. Int. Ed.* **47(40),** 7626–7630 (2008).

7. Ahmad, M., Gu, W., Geyer, T. & Helms, V. Adhesive water networks facilitate binding of protein interfaces. *Nat. Commun.* **2(261),** 1–7 (2011).

8. Plattner, N., Doerr, S., De Fabritiis, G. & Noé, F. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **9(10),** 1005–1011 (2017).

9. Blöchliger, N., Xu, M. & Caflisch, A. Peptide binding to a PDZ domain by electrostatic steering via nonnative salt bridges. *Biophys. J.* **108(9),** 2362–2370 (2015).

10. London, N., Movshovitz-Attias, D. & Schueler-Furman, O. The structural basis of peptide-protein binding strategies. *Structure* **18(2),** 188–199 (2010).

11. Nguyen, P. H., Okamoto, Y. & Derreumaux, P. Communication: Simulated tempering with fast on-the-fly weight determination. *J. Chem. Phys.* **138(6),** 061102 (2013).

12. Beutler, T. C. et al. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Lett.* **222(6),** 529–539 (1994).

13. Shaw D. E., Grossman, J. P., Bank, J. A., Batson, B., Butts, J. A., Chao, J. C., Deneroff, M. M., Dror, R. O., Even, A., Fenton, C. H., Forte, A., Gagliardo, J., Gill, G., Greskamp, B., Ho, C. R., Ierardi, D. J., Iserovich, L., Kuskin, J. S., Larson, R. H., Layman, T., Lee, L.-S., Lerer, A. K., Li, C., Killebrew, D., Mackenzie, K. M., Mok, S. Y.-H., Moraes, M. A., Mueller, R. Nociolo, L. J. Peticolas, J. L., Quan, T., Ramot, D., Salmon, J. K., Scarpazza, D. P., Schafer, U. B., Siddique, N., Snyder, C. W., Spengler, J., Tang, P. T. P., Theobald, M., Toma, H., Towles, B., Vitale, B., Wang, S. C. & Young, C. Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis (SC14)* 41–53 (IEEE, 2014).

14. McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L. P., Lane, T. J. & Pande, V. S. MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109(8),** 1528–1532 (2015).

15. Oliphant, T. E. Python for scientific computing. *Comput. Sci. Eng.* **9(3),** 10–20 (2007).

16. Theobald, D. L. & Wuttke, D. S. THESEUS: Maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics* **22(17),** 2171–2172 (2006).

17. Humphrey, W., Dalke, A. & Schulten, K. VMD—Visual Molecular Dynamics. *J. Mol. Graphics* **14(1),** 33–38 (1996).

18. The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. http://www.pymol.org.

19. Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **22(2),** 245–268 (1976).

20. Gumbart, J. C., Roux, B. & Chipot, C. Efficient determination of protein-protein standard binding free energies from first principles. *J. Chem. Theory Comput.* **9(8),** 3789–3798 (2013).

21. Wang, L., Friesner, R. A. & Berne, B. J. Replica exchange with solute scaling: A more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B* **115(30),** 9431–9438 (2011).

22. Schreiber, G. & Fersht, A. R. Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. *Biochemistry* **32(19),** 5145–5150 (1993).

23. Abriata, L. A. & Dal Peraro, M. Assessing the potential of atomistic molecular dynamics simulations to probe reversible protein-protein recognition and binding. *Sci. Rep.* **5,** 10549 (2015).

24. Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B* **119(16),** 5113–5123 (2015).

25. Dellago, C., Bolhuis, P. G. & Geissler, P. L. Transition path sampling. *Adv. Chem. Phys.* **123(1),** 1–81 (2002).

26. Peters, B. & Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.* **125(5),** 054108 (2006).

27. Best, R. B., Hummer, G. & Eaton, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U. S. A.* **110(44),** 17874–17879 (2013).

28. Kozakov, D. et al. The ClusPro web server for protein-protein docking. *Nat. Protoc.* **12(2),** 255–278 (2017).

29. Basu, S. & Wallner, B. DockQ: A quality measure for protein-protein docking models. *PLOS One* **11(8),** e0161879 (2016).

Supporting Tables and Figures

Complex (PDB ID)	Monomer 1 (PDB ID)	Interface RMSD (Å)	Monomer 2 (PDB ID)	Interface RMSD (Å)
Barnase-Barstar (1BRS)	Barnase (1A2P)	0.58	Barstar (1A19)	0.35
TYK2-Pseudokinase (4OLI)	TYK2 (3NZ0)	1.3	Pseudokinase (3ZON)	0.67
Ras–Raf-RBD (4G0N)	Ras (4RSG)	0.35	Raf-RBD (1RFA)	1.3
Insulin dimer (4INS)	Insulin monomer (2JV1)	1.3	N/A*	N/A*
RNaseHI–SSB-Ct (4Z0U)	Ribonuclease HI (2RN2)	2.0	N/A**	N/A**
CLC-ec1 dimer (1OTS)	CLC-ec1 monomer (3NMO)	1.0	N/A*	N/A*

*These systems are homodimers

**The SSB-Ct peptide is disordered in solution when not bound

Table S1. RMSD values of the binding interfaces of the protein-protein complexes. For the protein-protein complexes studied in this work, the RMSD between the binding interfaces of the experimentally determined associated and monomer structures is at most 2 Å. The protein-protein interface is defined as any pair of Cα atoms, one from each protein monomer, within 10 Å of each other.

Simulation	*System*	*Length (μs)*
BB_0_0	Barnase-Barstar	400
BB_0_1		400
BB_0_2		400
IND_0_0	Insulin dimer	140.7
IND_0_1		140.7
IND_0_2		140.7
IND_1_0		484.2
IND_1_1		484.2
IND_1_2		484.2
RAS_0_0	Ras–Raf-RBD	368.8
RAS_0_1		368.8
RAS_0_2		368.8
RNA0_0	RnaseHI–SSB-Ct	385.6
RNA0_1		385.6
RNA0_2		385.6
TYK_0_0	TYK2-Pseudokinase	218.1
TYK_0_1		218.1
TYK_0_2		218.1
TYK_1_0		212.1
TYK_1_1		212.1
TYK_1_2		212.1
CLC_0	CLC-ec1 Δ17–30 dimer	35.3
CLC_1		204.7

Table S2. List of tempered binding MD simulations. Tempered binding simulation parameters are described in the SI Appendix text. An aggregate simulation time of 6.9 ms was used for these simulations, with single simulations as long as 484 μs.

System	Tempered binding MD I-RMSD (Å)	Docking *Best predicted* I-RMSD (Å)	Docking *Best out of top 10* I-RMSD (Å)
Barnase-Barstar	0.61	1.1	1.1
Insulin dimer	0.76	12.1	1.8
Ras–Raf-RBD	1.3	10.7	1.8
TYK2-Pseudokinase	1.3	19.4	16.0
RNase HI–SSB-Ct	0.61	10.2	2.0

Table S3. A comparison between a representative structure of the thermodynamically most stable complex suggested by tempered binding and docked complexes predicted from an automated protein-protein docking webserver (ClusPro 2.0, accessed January, 2017[28]). The Cα I-RMSDs in this table were calculated using the DockQ program.[29] The purpose of this comparison was to obtain a rough sense of the accuracy of our tempered binding results compared to those from docking for the set of proteins studied in this work. It is possible that an expert user of ClusPro could achieve more accurate docking results than an automated webserver.

P2	**P3**	**P2'**	**P3'**
Barnase		*Barstar*	
70 to 73 89 to 92	85 to 88 97 to 100	13 to 24	66 to 79
Insulin		*Insulin*	
chain A: 13 to 19	chain B: 9 to 18	chain C: 13 to 19	chain D: 9 to 18
Ras		*Raf-RBD*	
52 to 57	17 to 25	83 to 88	57 to 62
TYK2		*Pseudokinase*	
942 to 953	986 to 9920	717 to 728	656 to 663
RNaseHI		*SSB-Ct*	
47 to 58	21 to 27	174	177

Table S4. Protein residue selections for defining the protein-protein orientation angles used in the angular distributions in SI Appendix Fig. S5. Following ref. 20, only the carbonyl carbon of each residue was selected when determining the center of mass; for all systems, P1 and P1' comprised all residues of the large and small monomers, respectively.

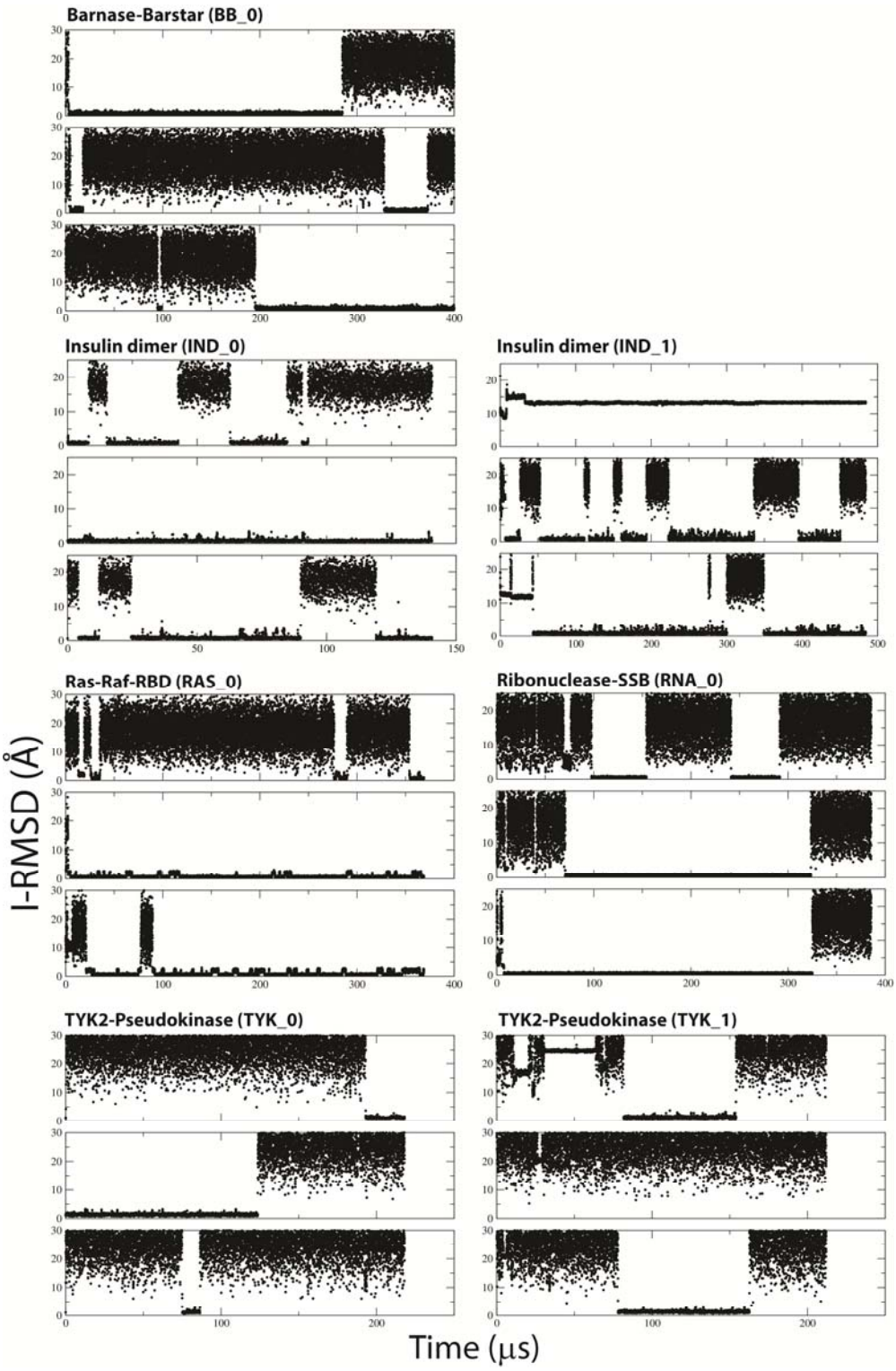

Figure S1. I-RMSD traces of the tempered binding MD simulations show reversible

association. See SI Appendix Table S1 for simulation codes. All simulations started from an unbound state, except for IND_0 and TYK_0, which started from the native-complex state.

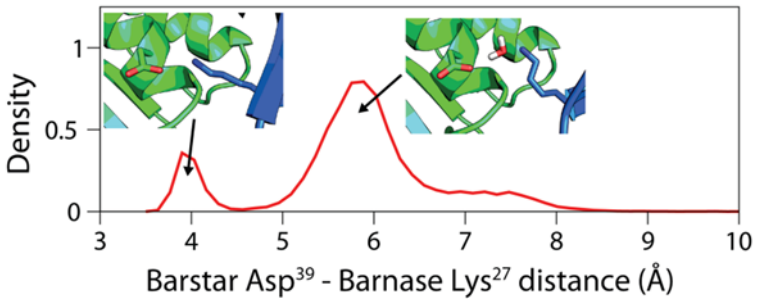

Figure S2. Tempered binding provides a direct, atomic-level observation of the ensemble of bound states involved in protein-protein interactions. For the barnase-barstar complex, the tempered binding simulations (BB_0) revealed fluctuations in the interaction between Lys^{27} on barnase and Asp^{39} on barstar between the weak, water-mediated interaction seen in the crystal structure, and a direct salt-bridge interaction only 1.1 kcal mol^{-1} higher in free energy, possibly explaining the stronger than expected interaction free energy seen in double-mutant cycle experiments for this pair of residues. A normalized density distribution of the distance between the Lys^{27} Nζ atom on barnase and the Asp^{39} Cγ atom on barstar in rung 0 of the tempered binding simulations of barnase-barstar shows two peaks corresponding to the water-mediated and direct salt-bridge interactions. The inset shows representative snapshots of the two states between which the system transitioned once bound.

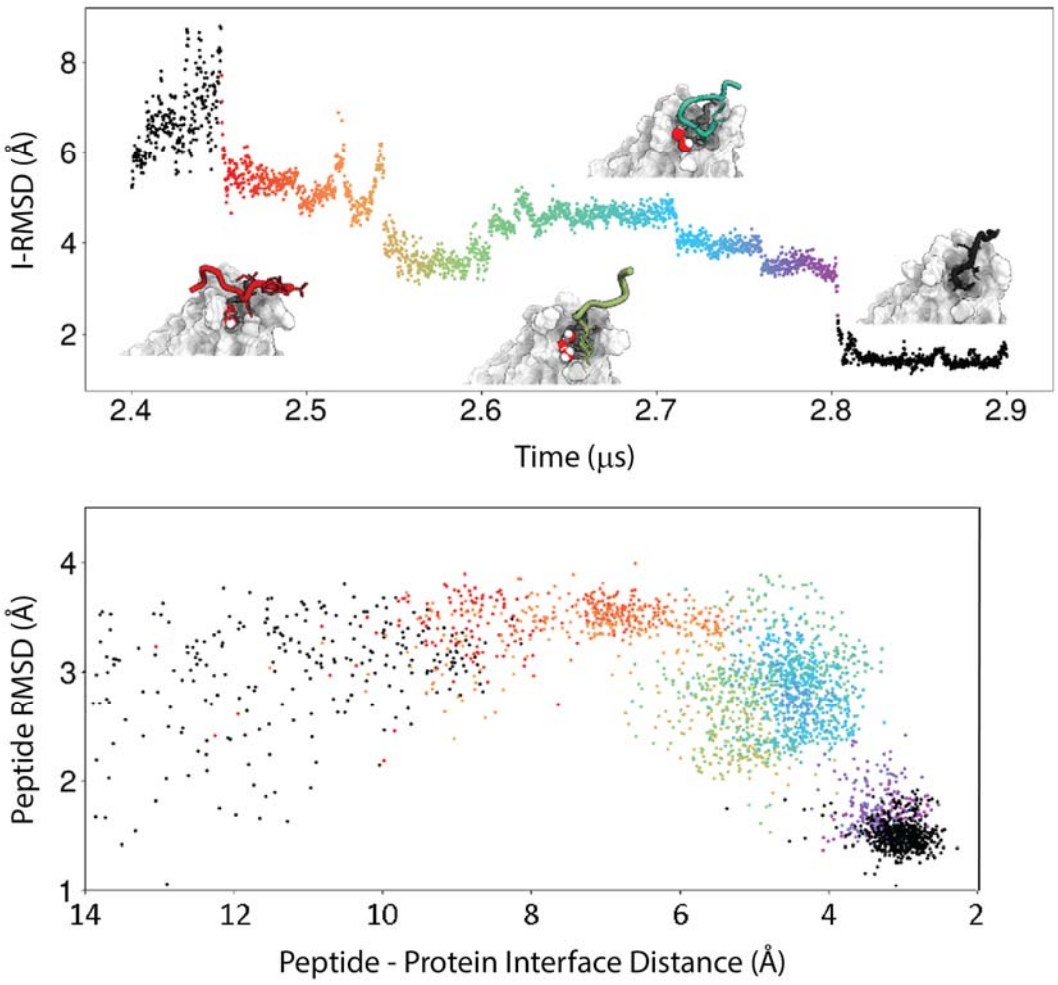

Figure S3. The SSB-Ct peptide folds upon binding to RNase HI. (Top) Heavy-atom I-RMSD as a function of time during a conventional MD simulation of SSB-Ct binding spontaneously to RNase HI. The first inset shows the initial contact between the binding region of the peptide (residues 174–177) and the RNase, which occurs at 2.45 μs (the SSB-Ct peptide is shown as a tube representation, with colors that correspond to the data points of the time trace, and the RNase HI is shown as a white surface). As the binding progresses (second and third

insets), the peptide partially inserts into the pocket and fluctuates among different conformations. The final inset shows the peptide achieving its native binding conformation. (Bottom) Peptide heavy atom RMSD (the "folding" coordinate) versus the distance between the centers of mass of the RNase HI's interface and the peptide (the "binding" coordinate). Colors correspond to the times shown in the top panel. The peptide gets close to the binding pocket before it finally folds into the correct conformation. Similar observations were made in other RNase HI–SSB-Ct binding trajectories.

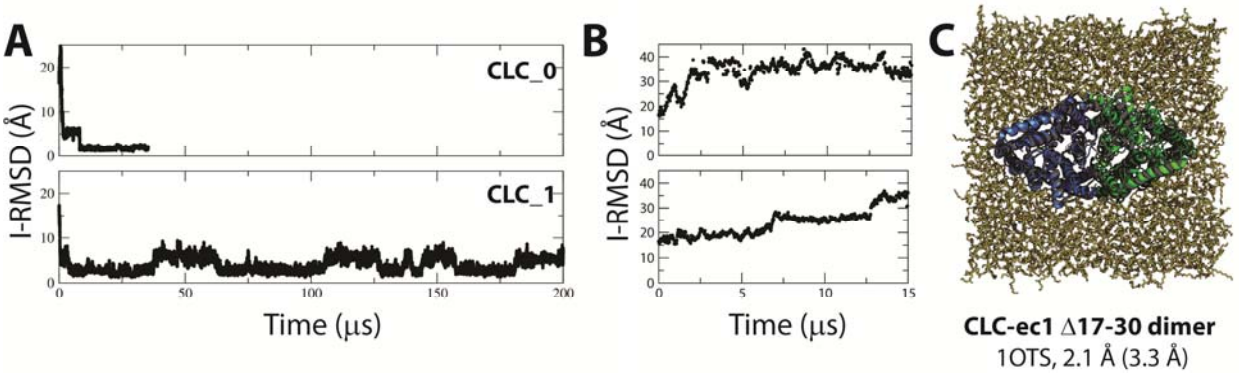

Figure S4. Tempered binding allows association of CLC-ec1 Δ17–30 to its experimentally determined dimer structure, but is unable to achieve reversible association. (A) I-RMSD traces of tempered binding simulations of CLC association at lower (CLC_0) and higher (CLC_1) scalings showing association to the experimentally determined dimer structure. (B) Association simulations of 10–15 μs using conventional MD become trapped in metastable states with I-RMSD values as high as 30–40 Å. (C) Overlay and comparison of a representative thermodynamically stable dimer structure from tempered binding to the crystal structure, determined as in Fig. 1.

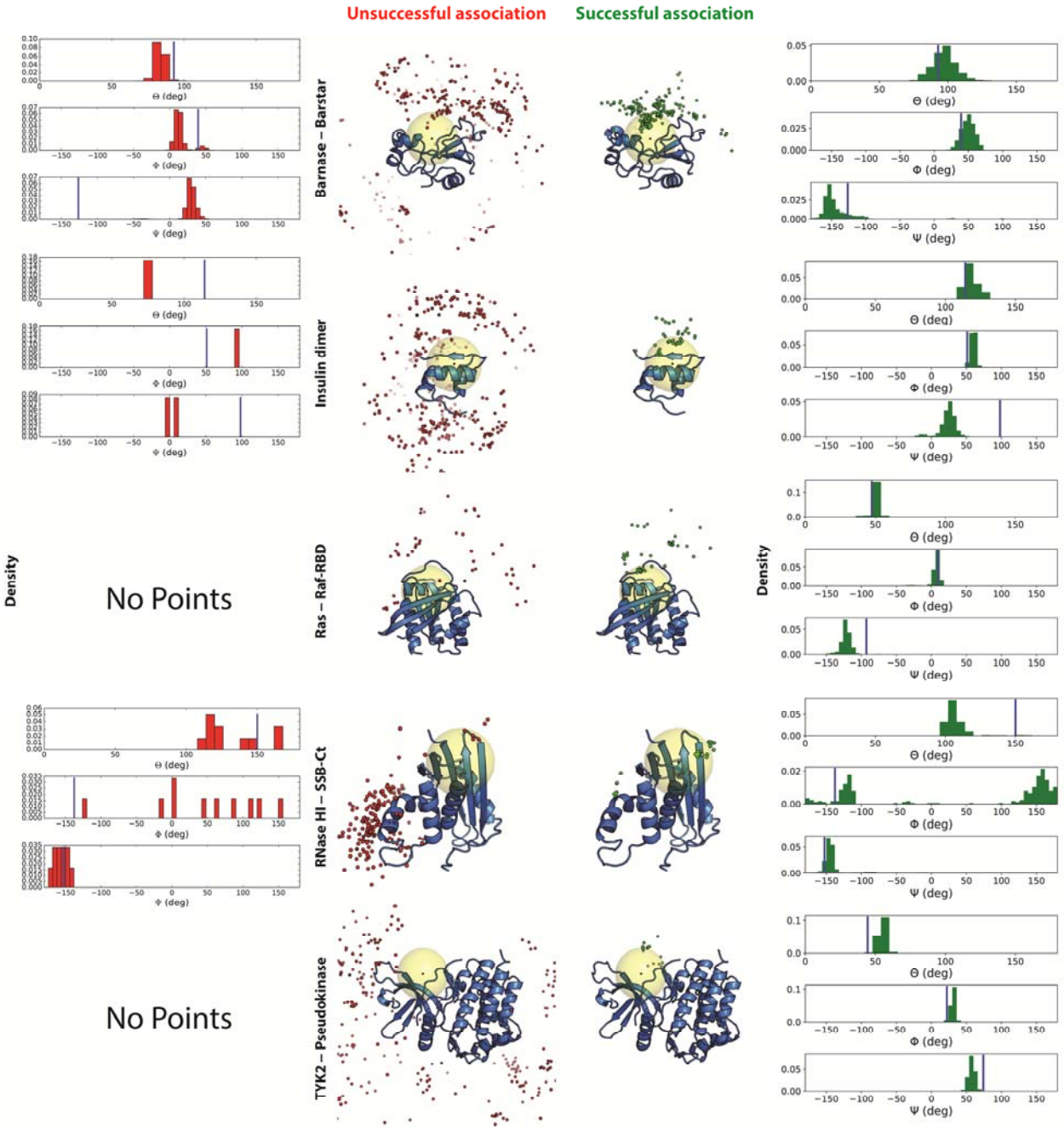

Figure S5. A moderate pre-orientation effect is seen in encounter-complex configurations during successful association. The two middle columns depict unsuccessful and successful association events as in Fig. 3. The plots in the left (red bars) and right (green bars) columns

show the angular distributions of the orientation of the smaller protein relative to the larger protein during unsuccessful and successful association trajectories, respectively, within the 10-Å yellow region. For Ras–Raf-RBD and TYK2–pseudokinase, no points were found in the yellow region for unsuccessful association, so no distributions are plotted. Vertical blue lines indicate the value of each angle in the experimentally determined structure. The definition of the orientation angles for each system is given in SI Appendix Table S4.

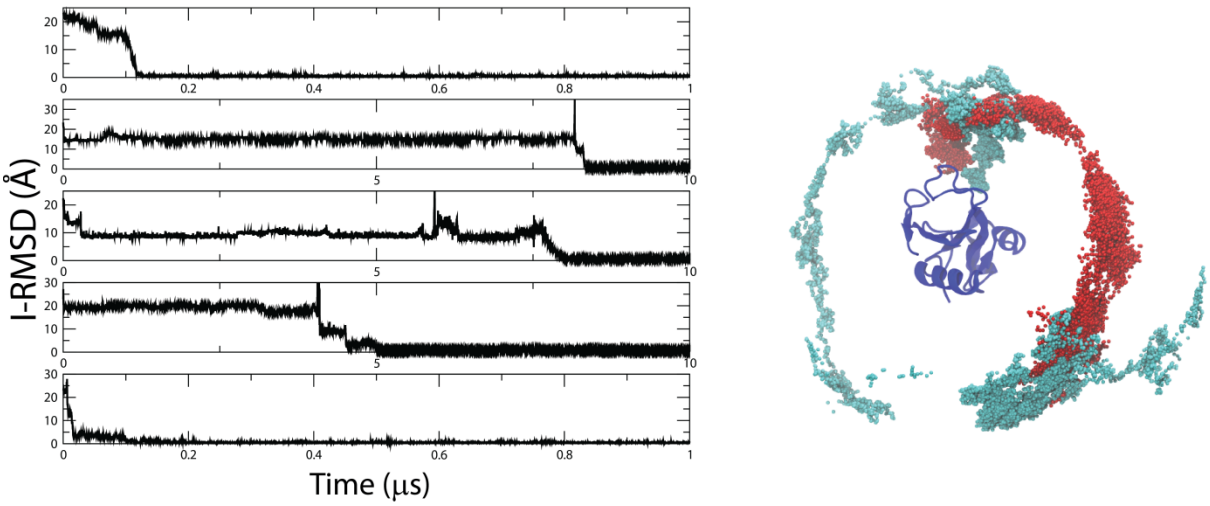

Figure S6. Spontaneous protein-protein association pathways explored different pathways prior to successful association. (Left) The interface RMSD traces of five spontaneous association trajectories of barnase-barstar, starting from the same unbound frame with different initial velocities. Large variations exist between different trajectories in terms of the length of time required for association and the behavior of the proteins prior to successful association. (Right) Two examples of successful association trajectories (red and cyan) in barnase-barstar sample different pathways prior to successful association. A single snapshot of the larger protein, barnase (blue cartoon), is shown for reference, overlaid with multiple snapshots of a Cα atom near the center of the native binding interface of the smaller protein, barstar (as in Fig. 3). In the cyan trajectory, the discontinuity between the two groups of points is due to a periodic boundary being crossed by barstar during association.

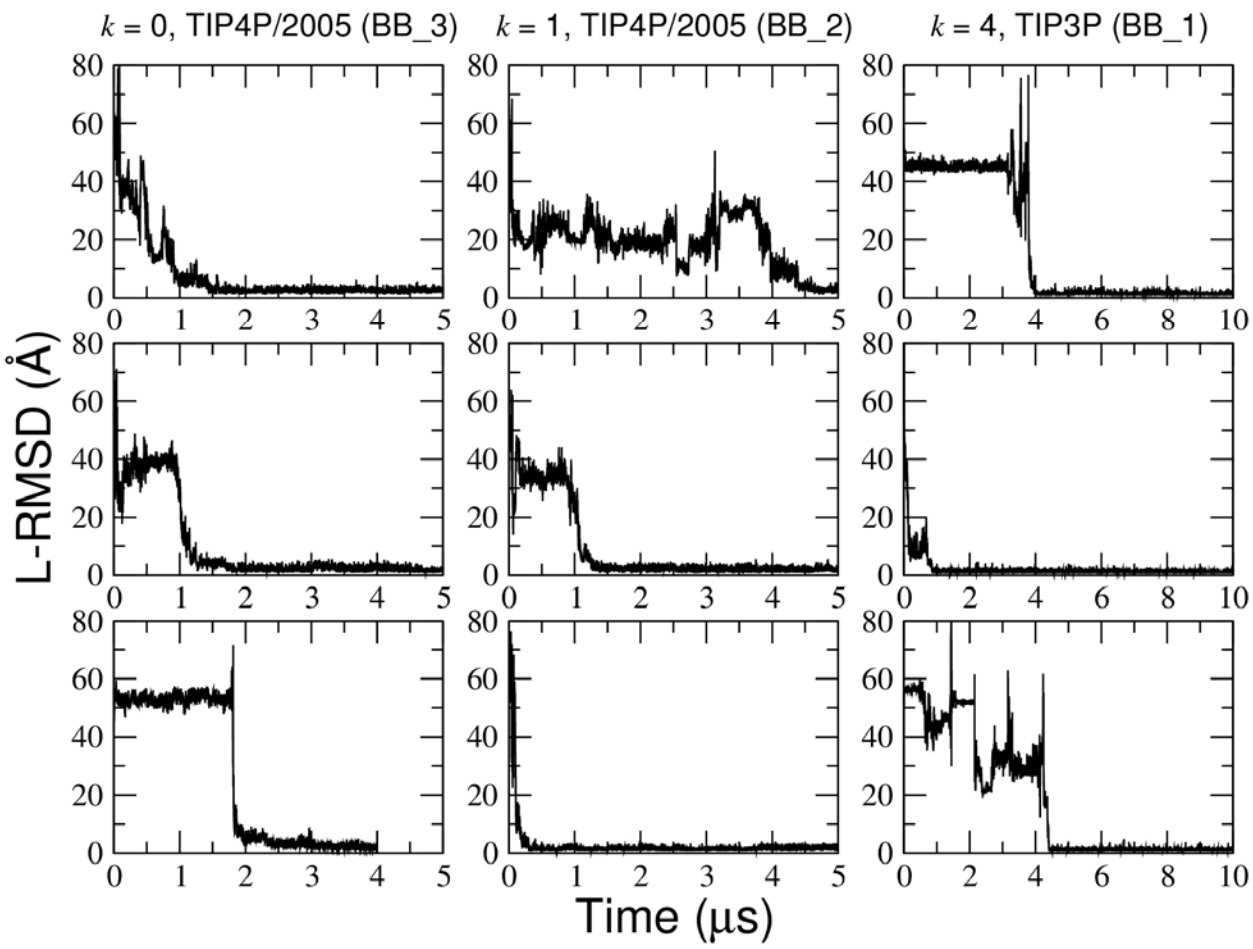

Figure S7. Spontaneous protein-protein association pathways for barnase-barstar in different force field conditions are qualitatively similar. Examples of L-RMSD traces of spontaneous association trajectories of barnase-barstar from three different force field conditions (BB_1, BB_2, and BB_3; see Table 1 and SI Appendix), in which there were differences in the strength of the torsional correction k (in units of kcal mol^{-1}) and the water model used.